**Big Data Programming**

**Team 4 Increment 1**

**Title:**

Twitter Web series Data Visualization with Spark ETL

**Team Members:**

* Avinash Ganguri
* Akhil Teja Kanugolu
* Geetanjali Makineni
* Bhashitha Siddareddy

**Introduction:**

Twitter Data Analysis with Spark ETL and visualizing is our title. Here we are taking the Twitter’s Web series data. The main objective of our project is doing the ETL process using the Spark’s Batch Processing and then Spark Integration using Web UI. The main source of our data is twitter and then collecting the data with Spark Batch Process. We can perform our transactions on the set of RDD’s and later we load our data in our Hive which is similarly equal to the ETL basic process. This is because we are living in a world where data handling and data using plays one most important role in making the decisions for most of the industries

**Background:**

1. We first collect the tweets from the Twitter api.

2. Later, we import the data that is collected into the hive from the HDFS.

3. Next we will export this data from there into RDBMS by usage of sqoop.

4. We later do the sentimental analysis on the tweets that are collected.

5. Now, we are using spark sql for writing required queries later by visualizing (examples like bar graphs and pie charts and some other type of graphs) the obtained results by using panda or Tableau.

**Goals and Objectives**

**Motivation:**

The motivation behind doing this analysis is we are living in a world where data handling and data using plays one most important role in making the decisions for most of the industries like Banking, Financial, Telecom and Health and IT sector serving ones. The main factors would be for getting the insights would be like managing its sheer volumes of data and its insights. Using the Apache Spark is one of the best amazing kind frameworks which will be handling big data and its real time performance of these analysis.

**Objectives:**

The main objective of our project is doing the ETL process using the Spark’s Batch Processing and then Spark Integration using Web UI. The main source of our data is twitter and then collecting the data with Spark Batch Process. We can perform our transactions on the set of RDD’s and later we load our data in our Hive which is similarly equal to the ETL basic process.

**Features:**

The main feature of the project is to collect the Real timed tweets from the twitter steaming API, also by performing the ETL which means we preprocess the data and extract the necessary data and then we load this data in our HIVE. Next, we use TextBlob for predicting the sentiment for each of the mentioned tweet. Later, we feed the data into our HDFS and then we implement SQL and Hive queries. Sqoop is used for transferring data between SQL and HDFS.

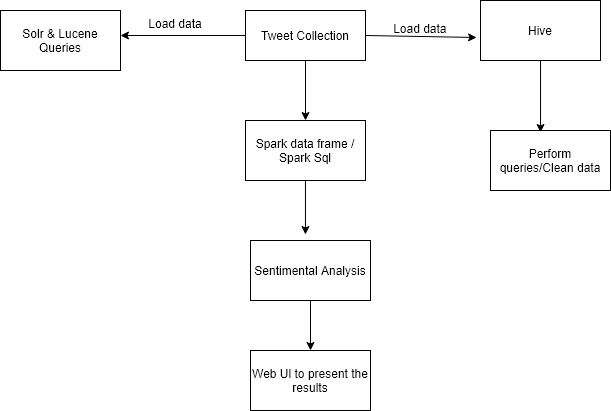
**Significance:**

Analyzing of sentimental analysis is done using one tool which is already existing named as ML tool which is TextBlob for prediction of sentiment on the tweets and Later we are using spark for writing the queries by visualizing with Panda.

**Dataset:**

We collect the data from Twitter with API with a developer account credentials. Here we took the keywords as Netflix, TV shows, etc., These tweets are in JSON file format. We have collected nearly like 2.5GB size. It has an information that shows like web series, tv series, documentary, Netflix shows, tv shows. We are collecting here using an API which is in batch format which is nearly downloading like around 5kb for a second where we filter the keywords like Netflix shows, TV shows etc.,

**Workflow:**



We have now first created one twitter developer account where next we got the tokens and credentials from API from where we took the twitter streaming and tweepy API in the python and stored tweets into the db. Extracted around 2.5 GB of dataset. Saved the data file in json format. Preprocess the data and create the database in spark framework with scala programming. Will be creating a data frame and use the Sql Context to execute the SQL queries. In our next phases we are implementing it on Solr and Lucene on our data and we also creating spark data frames on the file and creating different actions on data frames. We finally do the sentiment analysis on our data.

**Analysis:**

We have now first created one twitter developer account where next we got the tokens and credentials from API from where we took the twitter streaming and tweepy API in the python and stored tweets into the db. Extracted around 2.5 GB of dataset. Saved the data file in json format. Preprocess the data and create the database in spark framework with scala programming. Will be creating a data frame and use the Sql Context to execute the SQL queries.

**Implementation:**

**Tweets collected:**

* First, we got a developer access from the Twitter.After that we have created one app to use those API’s.
* The code for which twitter streaming and tweepy for collected tweets is below:

Text

Description automatically generated

Text

Description automatically generated

**Results:**

Starting the Spark-shell,

**Text

Description automatically generated**

**Query 1:**

Number of Tweets that are from different countries on web series.

Text

Description automatically generated

**Query 2:**

We are here fetching the count of tweets based on the co-ordinates of web series.

We got the information based on the longitude and latitude.

Text

Description automatically generated

**Query 3:**

To fetch the different languages in which the tweets are made about web series.

We got the count of number of tweets about the web series in separate languages.

Text

Description automatically generated

**Query 4:**

Fetching the users account names with a greater number of tweets on web series.

Text

Description automatically generated

**Query 5:**

Fetching the users that have more followers who are tweeting based on web series.

Text

Description automatically generated

**Query 6:**

Fetching the number of tweets based on different web series.

Text

Description automatically generated

**Query 7:**

Fetching the non-verified users who are having more number of followers.

Text

Description automatically generated

**Query 8:**

Users with a greater number of retweets for their tweets on web series.

Text

Description automatically generated

**Query 9:**

Number of Tweets that are on basis of different locations in the United States on web series.

Text

Description automatically generated

**Query 10:**

Fetching the top tweeted text and checking if someone re tweeted it on web series.

Text

Description automatically generated

**Project Management:**

Implementation Status Report:

Work Completed:

* Documenting and Downloading the data batch from twitter (Avinash, Geetanjali)
* Data Implementation & Data preprocessing along with documenting it (Akhil, Bhashitha)
* Queries executing and report of that (Avinash, Akhil, Geetanjali, Bhashitha)

Work to be Completed:

* Spark Sql Queries, Data Frames, and visualizations.
* Sentimental Analyzation (TextBlob)
* Solr Queries and Graph frames on the twitter data.
* Hosting on a Web Application.

**STORY TELLING:**

**WHO:**

The dataset is about the web series data where we sampled the tv shows, web series, Netflix series. The information disclosed about the shows watching around the world. Helps the advertiser’s for pushing the advertisement and series makers for making quick analysis.

**WHAT:**

The events like which show is being watched more an where it is being watched and which kind of people are watching are seen. It helps advertisers in pulling the data helps them in predicting the TRP which in result helps the series makers and advertisers.

**WHEN:**

The events take place on how people react to the series they tweet. As a result of COVID-19 most of the people settled back in their homes lacking entertainment.

**WHERE:**

All over the world. As everyone in the world haves an interest on web series and this data is being used everywhere. Thus, we are working on this data taking this into consideration.

**WHY:**

Most of the people believe in IMDB rating rather than genre concept. That is the reason behind doing this web series data collection.

**References:**

1. [developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json](https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json)
2. [www.info.com/**Twitter Data Visualization**](https://www.info.com/serp?q=twitter%20data%20visualization&segment=info.0376&s1aid=144007235&s1cid=357227086&s1agid=1264438724330452&s1kid=kwd-79027647541819:loc-190&utm_source=adcenter&msclkid=fa5513571bee1e925c7b4c24b6d1fd75)
3. [spark.apache.org/docs/latest/](https://spark.apache.org/docs/latest/streaming-programming-guide.html)